Chris Piech
CS109

Fall 2023
Oct 30th

CS109 Midterm Exam

This is a closed calculator/computer exam. You are, however, allowed to use notes in the exam. You have 2 hours (120 minutes) to take the exam. The exam is 120 points, meant to roughly correspond to one point per minute of the exam. You may want to use the point allocation for each problem as an indicator for pacing yourself on the exam.

In the event of an incorrect answer, any explanation you provide of how you obtained your answer can potentially allow us to give you partial credit for a problem. For example, describe the distributions and parameter values you used, where appropriate. It is fine for your answers to include summations, products, factorials, exponentials, and combinations. You can leave your answer in terms of Φ (the CDF of the standard normal). For example $\Phi(3/4)$ is an acceptable final answer..

I acknowledge and accept the letter and spirit of the honor code. I pledge to write more neatly than I have in my entire life:

Signature: _____

Family Name (print): _____

Given Name (print): _____

Stanford Email (@stanford.edu): _____

1 Measure of Variety [24 points]

In this question we are going to asks a simple question: what is the probability that two chosen objects from a set are different. This statistic is used both in Random Forest algorithms and in social science.

a. (5 points) Consider the following set of shapes. If you chose two shapes **with replacement** what is the probability that the two shapes are **the same**? Note that it is possible to get two triangles: after you pick the first triangle, you put it back into the set of shapes and it can be chosen again.

$$\begin{aligned} P(\text{same}) &= P(\text{two squares or two triangles}) \\ &= P(\text{two squares}) + P(\text{two triangles}) \\ &= P(\text{square})^2 + P(\text{triangle})^2 \\ &= \frac{1}{7}\cdot\frac{1}{7} + \frac{6}{7}\cdot\frac{6}{7} \\ &= \frac{37}{49} \end{aligned}$$

b. (5 points) Consider the following set of shapes. If you chose two shapes **with replacement** what is the probability that the two shapes are **different**?

Solution 1:

$$\begin{aligned} P(\text{different}) &= 1 - P(\text{same}) \\ &= 1 - (P(\text{two squares}) + P(\text{two triangles}) + P(\text{two stars})) \\ &= 1 - \left(\frac{4}{7}\cdot\frac{4}{7} + \frac{2}{7}\cdot\frac{2}{7} + \frac{1}{7}\cdot\frac{1}{7}\right) \\ &= 1 - \frac{3}{7} = \frac{4}{7} \end{aligned}$$

Solution 2:

$$\begin{aligned} P(\text{different}) &= P(\text{square then not square}) + P(\text{triangle then not triangle}) + P(\text{star then not star}) \\ &= \frac{4}{7}\left(1 - \frac{4}{7}\right) + \frac{2}{7}\left(1 - \frac{2}{7}\right) + \frac{1}{7}\left(1 - \frac{1}{7}\right) \\ &= \frac{4}{7}\cdot\frac{3}{7} + \frac{2}{7}\cdot\frac{5}{7} + \frac{1}{7}\cdot\frac{6}{7} \\ &= \frac{4}{7} \end{aligned}$$

c. (7 points) Consider a Poisson random variable $X \sim \text{Poi}(\lambda = 1)$. If you sample two numbers from X, what is the probability that the two numbers are **different**?

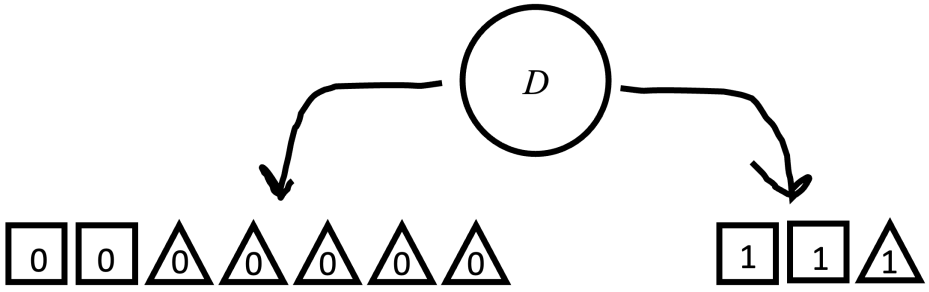
Solution 1:

$$\begin{aligned}
P(\text{different}) &= 1 - P(\text{same}) \\
&= 1 - \sum_{i=0}^{\infty} P(X = i)^2 \\
&= 1 - \sum_{i=0}^{\infty} \left[\frac{\lambda^i e^{-\lambda}}{i!}\right]^2 \\
&= 1 - \sum_{i=0}^{\infty} \left[\frac{e^{-1}}{i!}\right]^2 \\
&= 1 - \sum_{i=0}^{\infty} \frac{1}{(e \cdot i!)^2}
\end{aligned}$$

Solution 2:

$$\begin{aligned}
P(\text{different}) &= \sum_{i=0}^{\infty} P(X = i) \cdot P(X \neq i) \\
&= \sum_{i=0}^{\infty} P(X = i)(1 - P(X = i)) \\
&= \sum_{i=0}^{\infty} P(X = i) - \sum_{i=0}^{\infty} P(X = i)^2 \\
&= 1 - \sum_{i=0}^{\infty} P(X = i)^2 \\
&= 1 - \sum_{i=0}^{\infty} \left[\frac{\lambda^i e^{-\lambda}}{i!}\right]^2 \\
&= 1 - \sum_{i=0}^{\infty} \left[\frac{e^{-1}}{i!}\right]^2 \\
&= 1 - \sum_{i=0}^{\infty} \frac{1}{(e \cdot i!)^2}
\end{aligned}$$

d. (7 points) Each object in the decision tree below has **both** a value $\in \{0, 1\}$ and a shape.



A single decision node D sorts objects by their value. The Variety Statistic for an object with value v is the probability that, if you chose two objects with replacement from the set of objects with value v, the shapes of the chosen objects will be **the same**. An object is chosen at random from the 10 objects in the tree above. What is the expectation of the Variety Statistic of the chosen object?

The Variety Statistic (VS) depends on the value v. Uncertainty in the VS comes from the step where an object is chosen at random (from the 10). Once an object is chosen, v is known, and we can determine $VS(v)$. So to calculate expectation, we can consider the possible outcomes to be the VSs for the different possible values (here, 0 and 1).

$$\begin{aligned} VS(0) &= P(\text{same shape chosen from left set}) \\ &= P(\text{two squares or two triangles}) \\ &= \frac{2}{7} \cdot \frac{2}{7} + \frac{5}{7} \cdot \frac{5}{7} \\ &= \frac{29}{49} \end{aligned}$$

$$\begin{aligned} VS(1) &= P(\text{same shape chosen from right set}) \\ &= P(\text{two squares or two triangles}) \\ &= \frac{2}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{1}{3} \\ &= \frac{5}{9} \end{aligned}$$

$$\begin{aligned} E[VS] &= \sum_{v \in 0,1} VS(v) \cdot P(\text{value } v \text{ chosen}) \\ &= VS(0) \cdot P(\text{shape in left set chosen}) + VS(1) \cdot P(\text{shape in right set chosen}) \\ &= VS(0) \cdot \frac{7}{10} + VS(1) \cdot \frac{3}{10} \\ &= \frac{29}{49} \cdot \frac{7}{10} + \frac{5}{9} \cdot \frac{3}{10} \\ &= \frac{29}{70} + \frac{5}{30} \\ &= \frac{122}{210} \end{aligned}$$

2 GPT Generation [27 points]

GPT-4 is a popular A.I. model used to generate text. GPT-4 is based on a Large Language Model (LLM) that takes in a string of text, and quickly calculates the probability distribution over the next word (formally called a token). GPT-4 uses this next-word probability to generate complete responses.

You are given access to a function `gpt_pr(s)` that returns a dictionary with conditional probabilities. The keys of the dictionary are all the possible next words. For each key x, the corresponding value is the probability that x is the next word, given s has been produced so far. The length of the dictionary will always be 10^6 for the million most common words in English. A special key is `'</s>'` which indicates that the LLM should stop producing text.

For example, a call to `gpt_pr("")` returns the probabilities that a word would be the first word in a response. Calling `gpt_pr("I love")` returns the conditional probability of the 10^6 next words, given "I love" has already been produced in the response. For example `gpt_pr("I love")` could return:

```
{ "you":0.1, "CS109":0.01, "serendipity":0.001, ..., "</s>":0.001}
# where gpt_pr("I love")["you"] = P(next = "you" | so far = "I love")
```

One algorithm for producing a full response is "Most Likely Algorithm" (MLA). MLA is a surprisingly bad.

> **Def. Most Likely Algorithm (MLA)**: For each possible response with ≤ 100 words, calculate the probability of generating that response. Then return the response with the highest probability.

a. (6 points) Use `gpt_pr(s)` to compute the probability of generating the three words `"I love serendipity"`, if your starting point is the empty string ("").

Using the chain rule:

$$\begin{aligned} P(\text{“I love you”}) &= P(\text{next = “I” } | \text{ so far = “”}) \\ &\cdot P(\text{next = “love” } | \text{ so far = “I”}) \\ &\cdot P(\text{next = “you” } | \text{ so far = “I love”}) \\ &= \texttt{gpt\_pr("")["I"]} \\ &\cdot \texttt{gpt\_pr("I")["love"]} \\ &\cdot \texttt{gpt\_pr("I love")["serendipity"]} \end{aligned}$$

b. (8 points) Issue 1: MLA needs to test all possible complete productions with up to 100 words. If your starting point is the empty string, how many complete productions are there with ≤ 100 words? To be a complete string the last word **must** be `'</s>'` and no other word can be an `'</s>'`.

The number of complete productions with k words is $(10^6)^{k-1}$, since there are 10^6 options for each word except the last word, which must be the stop word. So the number of complete productions with up to 100 words is:

$$\sum_{i=0}^{99} 10^{6^i}$$

or

$$\sum_{i=0}^{99} (10^6 - 1)^i$$

c. (9 points) Issue 2: You observe that MLA prefers short strings. Consider MLA selecting a response starting from the empty string. Use a geometric random variable as a metaphor to explain why a production with just the stop word could be the most likely. Also note why the geometric is an imperfect model for the length of an LLM response.

Assume that at each word there is a constant, independent probability p that the stop word will be next. Let X be the words until the stop word. $X \sim \text{Geo}(p)$. Then the reason why the stop word would be chosen most often is that the geometric has highest probability for $X = 1$ regardless of p. This is an imperfect metaphor because the probability of the stop word is not independent of the prior words chosen so far, nor is the probability of choosing the stop word at each "trial" constant.

d. (4 points) Issue 3: Humans find MLA productions unnatural. One researcher claims the responses generated by MLA have too high a variance in *probabilities* for human taste. For a single production of length 100, you have a list of next-word probabilities $L = [p_1 \dots p_{100}]$ where $L[i]$ is the probability of generating word i, given words 0 to $i-1$. What is the formula you would use to calculate the variance of the probabilities in list L.

Solution 1:

$$\text{Var}(X) = \frac{1}{100}\sum_{i=1}^{100}(p_i - E[X])^2$$

where, by the definition of expectation, $E[X] = \frac{1}{100}\sum_{i=1}^{100} p_i$, or the average of the values in list L.

Solution 2:

$$\text{Var}(X) = E[X^2] - E[X]^2$$

$$E[X]^2 = \left(\frac{1}{100}\sum_{i=1}^{100} p_i\right)^2$$

$$E[X^2] = \frac{1}{100}\sum_{i=1}^{100} p_i^2$$

3 Era's Tour by Ticketmaster [24 points]

You have 1000 tickets to sell from an online website. You will open ticket sales at midnight. At midnight there are 1500 people online, waiting to buy tickets.

Each person takes X seconds to buy one ticket. Assume that for each person, on average, it takes them 20 seconds to purchase a ticket. People purchase tickets independently, and the rate of ticket purchasing is constant over time.

a. (8 points) What is the probability that a single individual buys their ticket within 20 seconds?

Let p_a be the probability that an individual checks out within 20 seconds.

Using $X \sim \text{Exp}(\lambda = 1/20)$, with a time unit of 1 second:

$$\begin{aligned} p_a &= P(X < 20) \\ &= 1 - e^{-\frac{20}{20}} \\ &= 1 - e^{-1} \end{aligned}$$

Other time units used with the exponential also work.

Using $X \sim \text{Poi}(\lambda = 1)$, with a time unit of 20 seconds:

$$\begin{aligned} p_a &= 1 - P(X = 0) \\ &= 1 - e^{-1} \end{aligned}$$

b. (8 points) Suppose it is midnight and ticket sales have just opened. What is the probability that there are still tickets after 20 seconds? Give a formula for the exact answer. Let $p_a = 0.63$ be your answer to part a.

Let Y be the number of people who buy tickets within 20 seconds. $Y \sim \text{Bin}(n = 1500, p_a)$. There are still tickets left if Y is less than 1000:

$$\begin{aligned} P(Y < 1000) &= \sum_{i=0}^{999} P(Y = i) \\ &= \sum_{i=0}^{999} \binom{1500}{i} p_a^i (1-p_a)^{1500-i} \end{aligned}$$

Here is another possible approach using complementary probability.

$$\begin{aligned} P(Y < 1000) &= 1 - P(Y >= 1000) \\ &= 1 - \sum_{i=1000}^{1500} P(Y = i) \\ &= 1 - \sum_{i=1000}^{1500} \binom{1500}{i} p_a^i (1-p_a)^{1500-i} \end{aligned}$$

A common misconception was to forget to include the summation and write $\binom{1500}{1000} p_a^{1000} (1-p_a)^{500}$ since there are only 1000 tickets. However, note that this is an incorrect complement since

$$\binom{1500}{1000} p_a^{1000} (1-p_a)^{500} + \sum_{i=0}^{999} \binom{1500}{i} p_a^i (1-p_a)^{1500-i} \neq 1$$

c. (8 points) Provide an approximation for your answer to part (b) that can be computed in constant time.

Let A be the approximating normal for Y. $A \sim N(\mu = 1500 \cdot p_a, \sigma^2 = 1500 \cdot p_a \cdot (1 - p_a)$

$$\begin{aligned} P(Y < 1000) &\approx P(A < 999.5) \\ &= \Phi\Big(\frac{999.5 - 1500 \cdot p_a}{\sqrt{1500 \cdot p_a \cdot (1 - p_a)}}\Big) \end{aligned}$$

4 How Many Coin Flips? [20 points]

A person is flipping a fair coin an unknown number of times. When the flipping is over, you are told that there were exactly 5 heads. Let N be a random variable for the number of times that the coin was flipped. Your prior belief is that any number of flips from 0 to 100 inclusive is equally likely. Provide **both** a math expression and pseudo code for $P(N = i)$ given that there were exactly 5 heads and i is between 0 and 100 inclusive.

This is an inference problem. We can apply Bayes' Rule:

$$P(N = n|5 \text{ heads}) = \frac{P(5 \text{ heads}|N = n)P(N = n)}{P(5 \text{ heads})}$$

$$P(N = n|5 \text{ heads}) = K \cdot P(5 \text{ heads}|N = n)P(N = n)$$

Because the denominator would be hard to calculate, we can re-write it as a constant K.

Our prior belief that all numbers of flips from 0 to 100 are equally likely means that for $n \in \{0, ..., 100\}$, $P(N = n) = \frac{1}{101}$. Given any number of total flips n, the number of heads can be represented as a binomial: $X \sim \text{Bin}(n, 0.5)$. So our likelihood is the binomial PMF evaluated at $X = 5$.

$$P(N = n|5 \text{ heads}) = K \cdot \binom{n}{5} 0.5^n \cdot \frac{1}{101}$$

Since we observe 5 heads, we know that at least 5 flips happened, so $P(N < 5) = 0$. Thus the possible values for n with nonzero probabilities are 5 through 100.

To find a value for K, we just need to ensure that the probabilities of all possible values for n will sum to 1, so we can set K to the inverse of that:

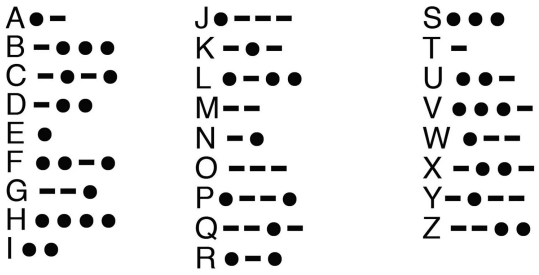
$$K = \frac{1}{\sum_{i=5}^{100} \binom{n}{5} 0.5^n \cdot \frac{1}{101}}$$

In pseudocode:

```
probs_list_all_n = []

for i in range(101): # loop through 0 to 100 inclusive
. if i < 5:
. . probs_list_all_n.append(0)

. else:
. . prior = 1 / 101
. . likelihood = math.comb(i, 5) * (0.5 ** i)
. . prior_times_likelihood = prior * likelihood
. . probs_list_all_n.append(prior_times_likelihood)

sum_of_all_probs = sum(probs_list_all_n) # Bayes denominator, or 1/K

return probs_list_all_n[n] / sum_of_all_probs
```

5 Auto Morse Code Detection [25 points]

We are building an auto-morse code detector. Morse code is a system of sending messages which are encoded as short-length sounds, "dots" and long-length sounds "dashes".



The duration of a dot sound is Gaussian with mean of 0.5 seconds and standard deviation of 0.3.
The duration of a dash sound is Gaussian with a mean of 1.5 seconds and standard deviation of 1.0.

a. (6 points) What is the probability that a dot sound will be longer than 1.0 second?

Let D be the duration of the sound. $D|\text{dot} \sim Normal(0.5, 0.3^2)$.

$$\begin{aligned} P(D|dot > 1) &= 1 - P(D|dot < 1) \\ &= 1 - P\left(Z < \frac{1-0.5}{0.3}\right) \\ &= 1 - \Phi\left(\frac{5}{3}\right) \end{aligned}$$

b. (8 points) Calculate the **prior** belief that a sound is a dot. Assume that messages are formed using the letters of the alphabet. For each letter, let $p_a, p_b, p_c, \ldots, p_y, p_z$ represent the likelihood of that letter appearing in a message. For any letter α, let dots(α) specify the count of dots and dashes(α) indicate the count of dashes in its morse code representation. As an example for g, $p_g = 0.02$, dots(g) = 1, dashes(g) = 2.

$$\begin{aligned} P(E) &= \sum_{i=a}^{z} P(D|L=i)P(L=i) \\ &= \sum_{i=a}^{z} \frac{\text{dots}(i)}{\text{dots}(i) + \text{dashes}(i)} \cdot p_i \end{aligned}$$

c. (8 points) Let p_{dot} be your answer to the previous part. A sound is observed with duration 1.1 seconds. What is the probability that it is a dot?

$$
\begin{aligned}
P(\text{dot}|D=1.1) &= \frac{P(D=1.1|\text{dot})\cdot P(\text{dot})}{P(D=1.1|\text{dot})\cdot P(\text{dot}) + P(D=1.1|\text{dash})\cdot P(\text{dash})} \\
&= \frac{f(D=1.1|\text{dot})\cdot p_{dot}}{f(D=1.1|\text{dot})\cdot p_{dot} + f(D=1.1|\text{dash})\cdot (1-p_{dot})} \\
&= \frac{\frac{1}{0.3\sqrt{2\pi}} e^{-\frac{(1.1-0.5)^2}{2\cdot 0.3^2}} \cdot p_{dot}}{\frac{1}{0.3\sqrt{2\pi}} e^{-\frac{(1.1-0.5)^2}{2\cdot 0.3^2}} \cdot p_{dot} + \frac{1}{1\sqrt{2\pi}} e^{-\frac{(1.1-1.5)^2}{2\cdot 1^2}} \cdot (1-p_{dot})} \\
&= \frac{\frac{10}{3} e^{-2} \cdot p_{dot}}{\frac{10}{3} e^{-2} \cdot p_{dot} + e^{-0.08} \cdot (1-p_{dot})}
\end{aligned}
$$

d. (3 points) There is a value, t, such that for any sound duration greater than t, the probability the sound is a dash is greater than the probability the sound is a dot. Calculate t. Note that this problem is intentionally given fewer points .

$$
\begin{aligned}
P(\text{dot}|D=t) &= P(\text{dash}|D=t) \\
\frac{P(D=t|\text{dot})P(\text{dot})}{P(D=t)} &= \frac{P(D=t|\text{dash})P(\text{dash})}{P(D=t)} \\
P(D=t|\text{dot})p_{\text{dot}} &= P(D=t|\text{dash})(1-p_{\text{dot}}) \\
\frac{P(D=t|\text{dot})}{P(D=t|\text{dash})} &= \frac{(1-p_{\text{dot}})}{p_{\text{dot}}} \\
\frac{\frac{1}{0.3} e^{-\frac{(t-0.5)^2}{0.18}}}{e^{-\frac{(t-1.5)^2}{2}}} &= \frac{(1-p_{\text{dot}})}{p_{\text{dot}}} \\
e^{\frac{(t-1.5)^2}{2} - \frac{(t-0.5)^2}{0.18}} &= \frac{0.3(1-p_{\text{dot}})}{p_{\text{dot}}} \\
\frac{(t-1.5)^2}{2} - \frac{(t-0.5)^2}{0.18} &= \log\left(\frac{0.3(1-p_{\text{dot}})}{p_{\text{dot}}}\right)
\end{aligned}
$$

That's all folks! We hope you had fun. Here are some optional notes for further curiosity.

i. The measure of variety you calculated is a fast and efficient replacement for information gain in classification. As such it has made its way into a ton of algorithms. In random forests, decision nodes are created which minimize the variety of class labels in each of the children.

ii. The best algorithm for producing responses using an LLM is an open problem. It has been observed that optimistic algorithms, such as Beam Search substantially outperform the Most Likely algorithm even when it is tractable. Why? It is a mystery! If you want to learn more, see the Meister et al. 2020 paper "If beam search is the answer, what was the question?"

iii. Ticketmaster hosted a disastrous sale of tickets for Taylor Swifts Era's tour. They say 3.5 million people had pre-registered for Taylor's Verified Fan sale, which was, it added, the "largest registration in history". They certainly need a new user experience, and someone who knows a thing or two about probability. Image was drawn by Dall-E with the prompt: "Minimalistic hand-drawn sketch of a female singer with wavy blonde hair, singing into a microphone, using light strokes and minimal ink."

iv. How would you have changed your answer if I provided you with two observations (say a conditionally independent observation of 7 heads)? This is not a midterm question, but it could have been!

v. Morse code has been adapted for use by individuals with certain disabilities, allowing them to communicate. For instance, some assistive communication devices use Morse code as an input method for those who have limited motor skills but can make consistent keystrokes.